Supplementary Materials

Title: "Inverse uncertain characteristics of pollution source identification for river chemical spill incidents by stochastic analysis"

Authors: Jiping Jiang*, Feng Han, Yi Zheng, Nannan Wang, Yixing Yuan

*Email: jiangjp@sustc.edu.cn

Contents:

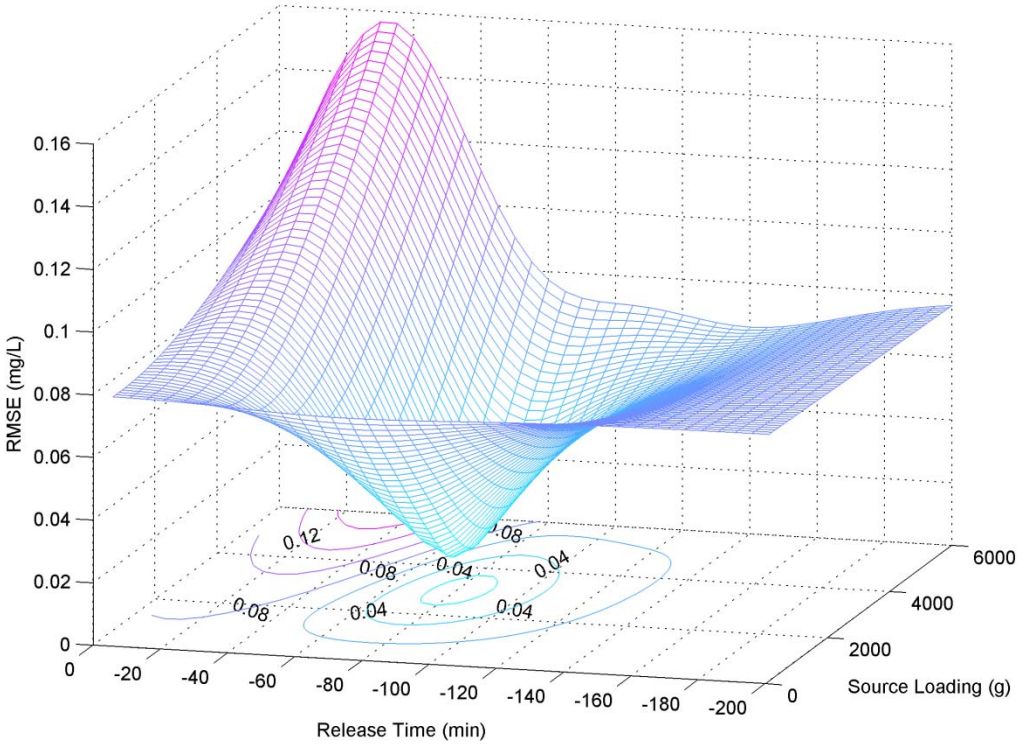


Fig.S1 Mesh plot of objective function (RMSE) for instantaneous discharge when source loading (M_s) and release time (t_s) are unknown.

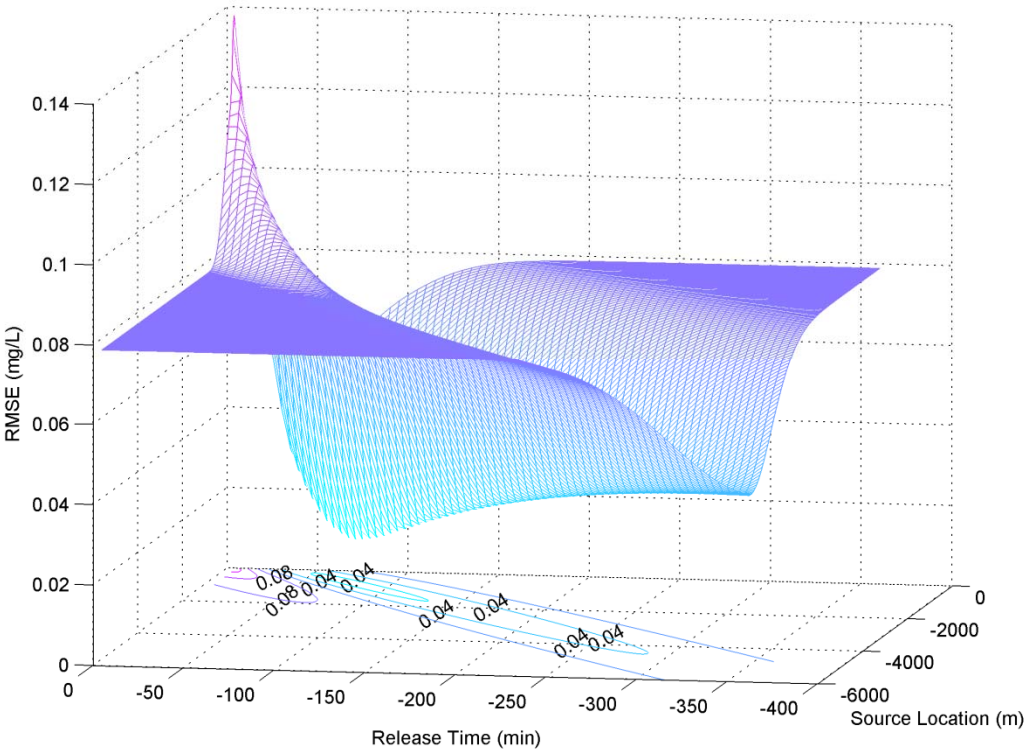


Fig. S2 Same as Fig. S1 but source location (x_s) and release time (t_s) are unknown.

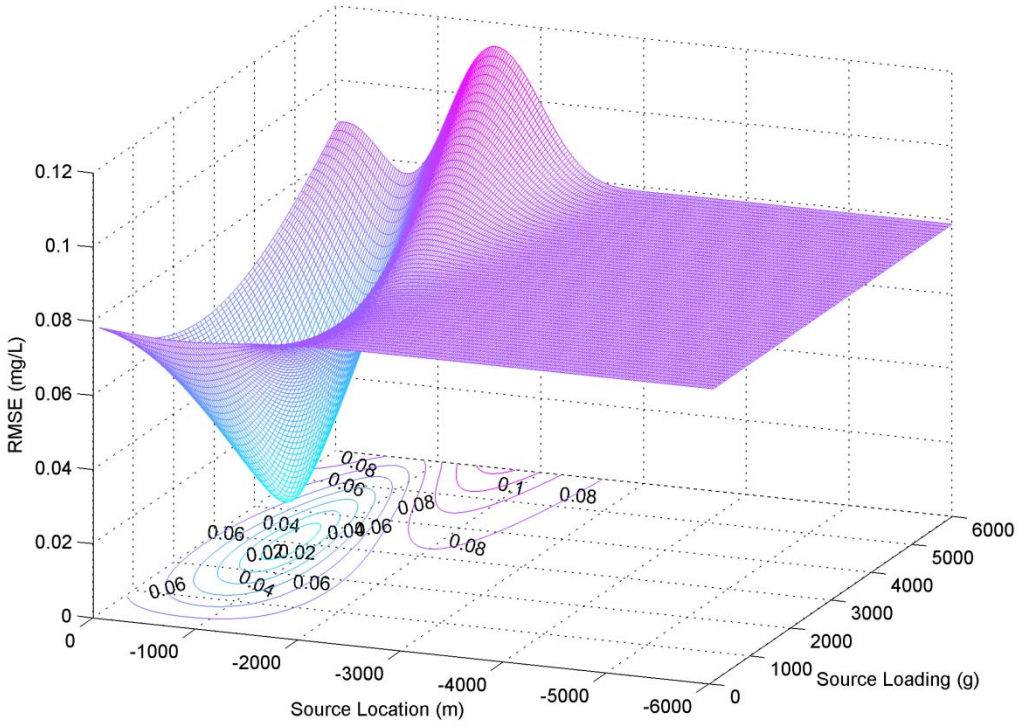


Fig. S3 Same as Fig. S1 but source loading (M_s) and source location (x_s) are unknown.

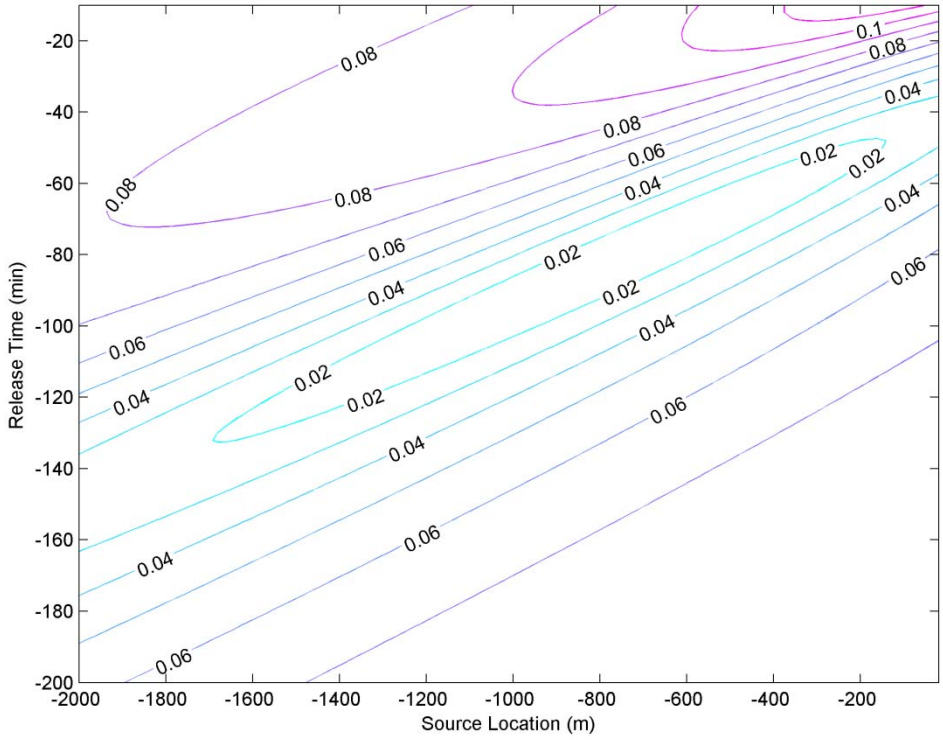


Fig. S4 Contour plot of objective function (RMSE) for instantaneous discharge when source location (x_s) and release time (t_s) are unknown.

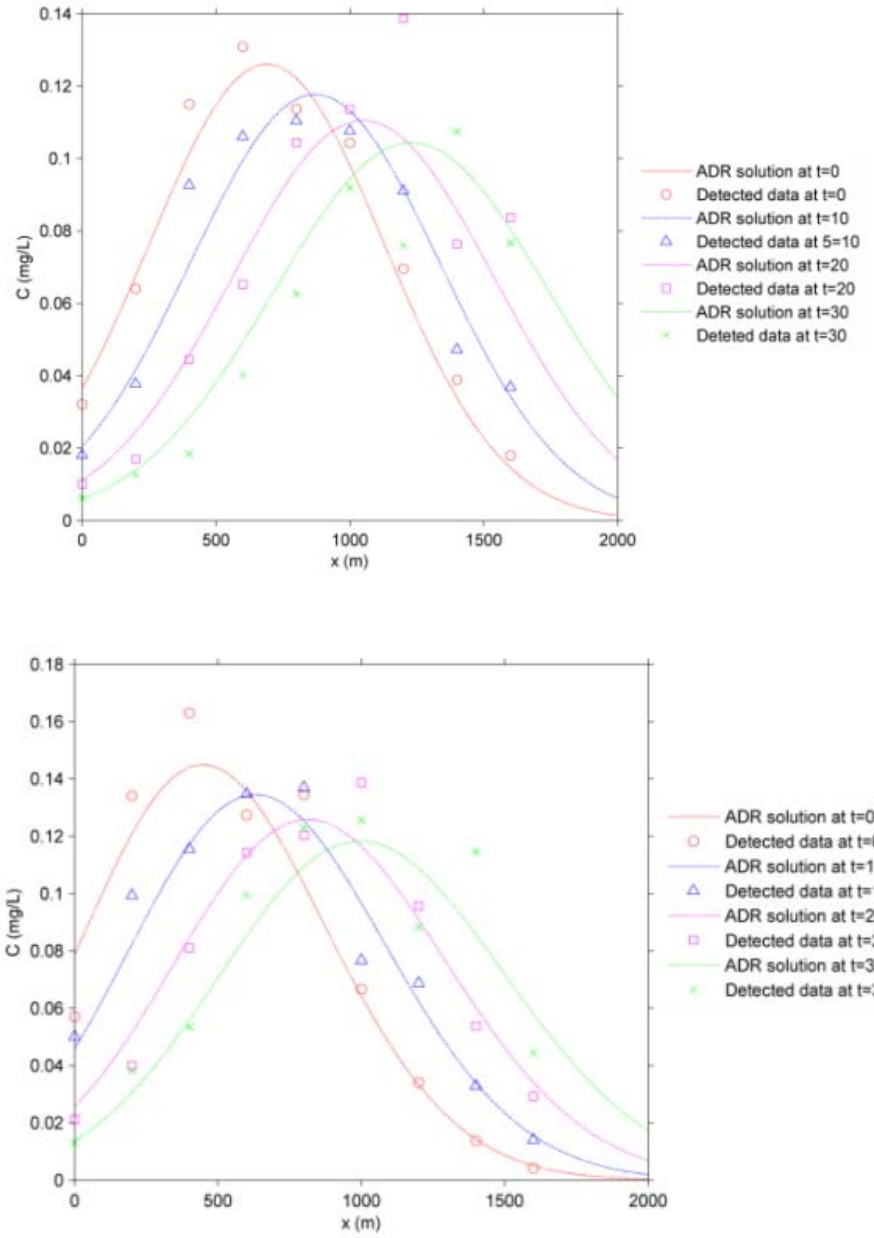


Fig. S5 Synthetic data for Case-S1(instantaneous discharge) and Case-S2 (continuous discharge)

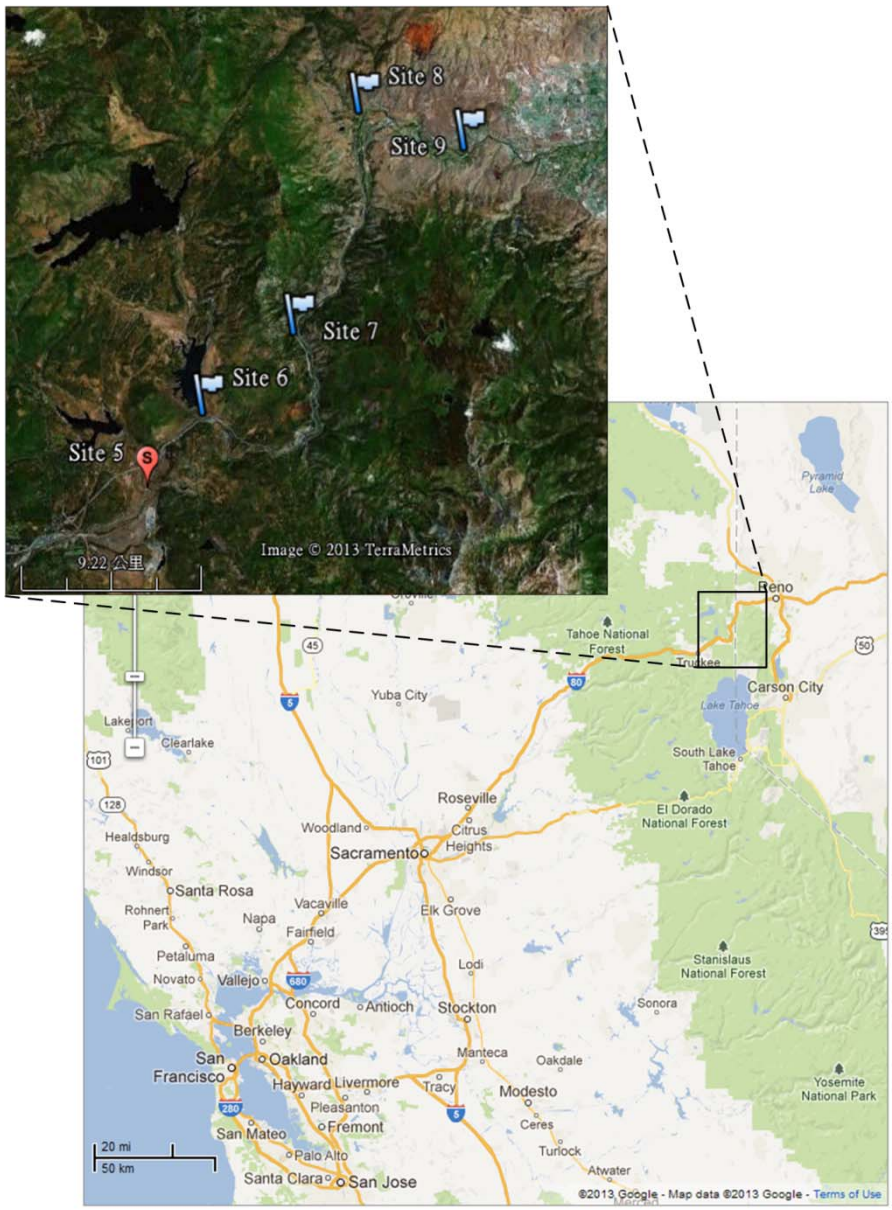


Fig. S6 Locations of Trackee River injection and sampling sites used in Case-T1. Site7-9 represents 13.74km, 22.64km, 27.4 km downstream the injection point respectively.

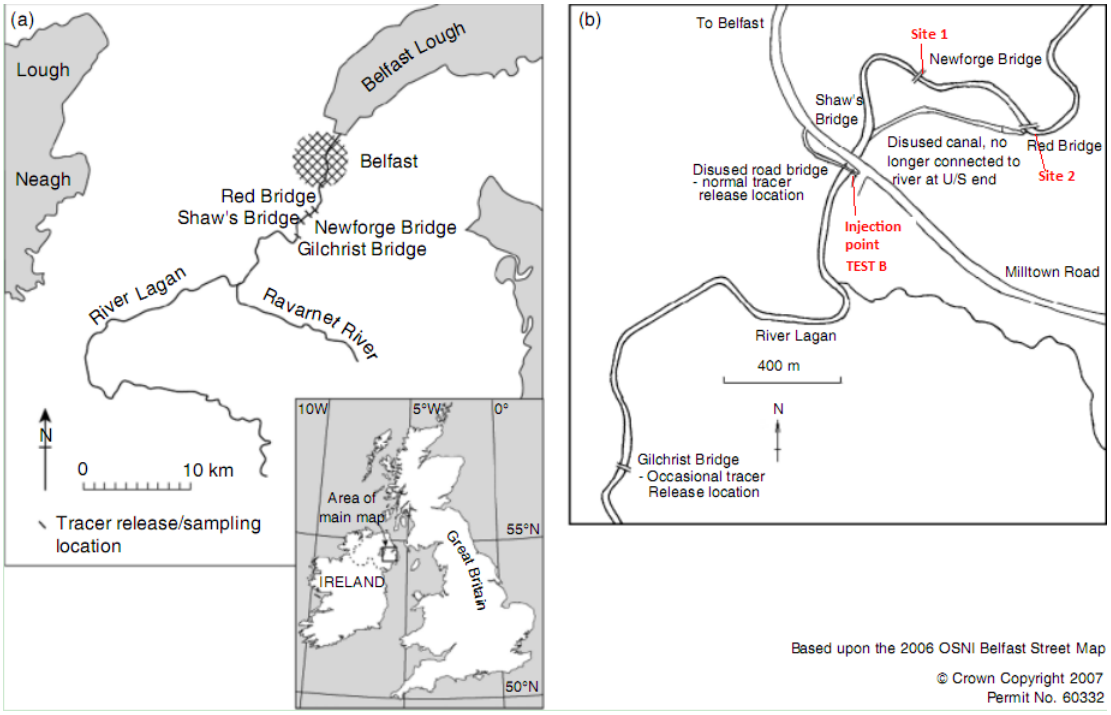


Fig. S7 Locations of River Lagan injection and sampling sites used in Case-T2. Site1 and Site 2 represents 600m and 1200m downstream the injection point respectively.

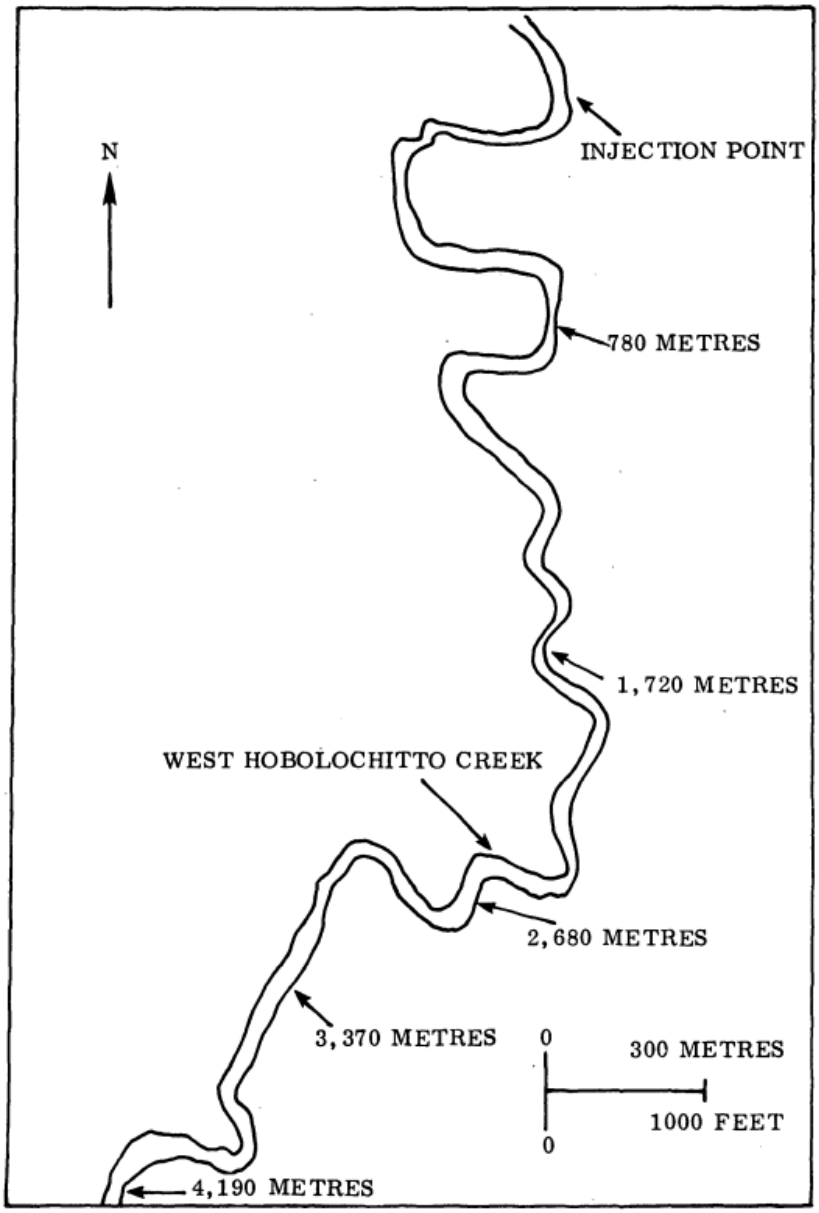


Fig. S8 Sketch of the experimental reach of West Hobolochitto Creek (Case-T3), showing the relative locations of the injection and sampling points (Rathbun,1975). Site 3-5 are 2680m, 3370m, and 4190m downstream from the injection point.

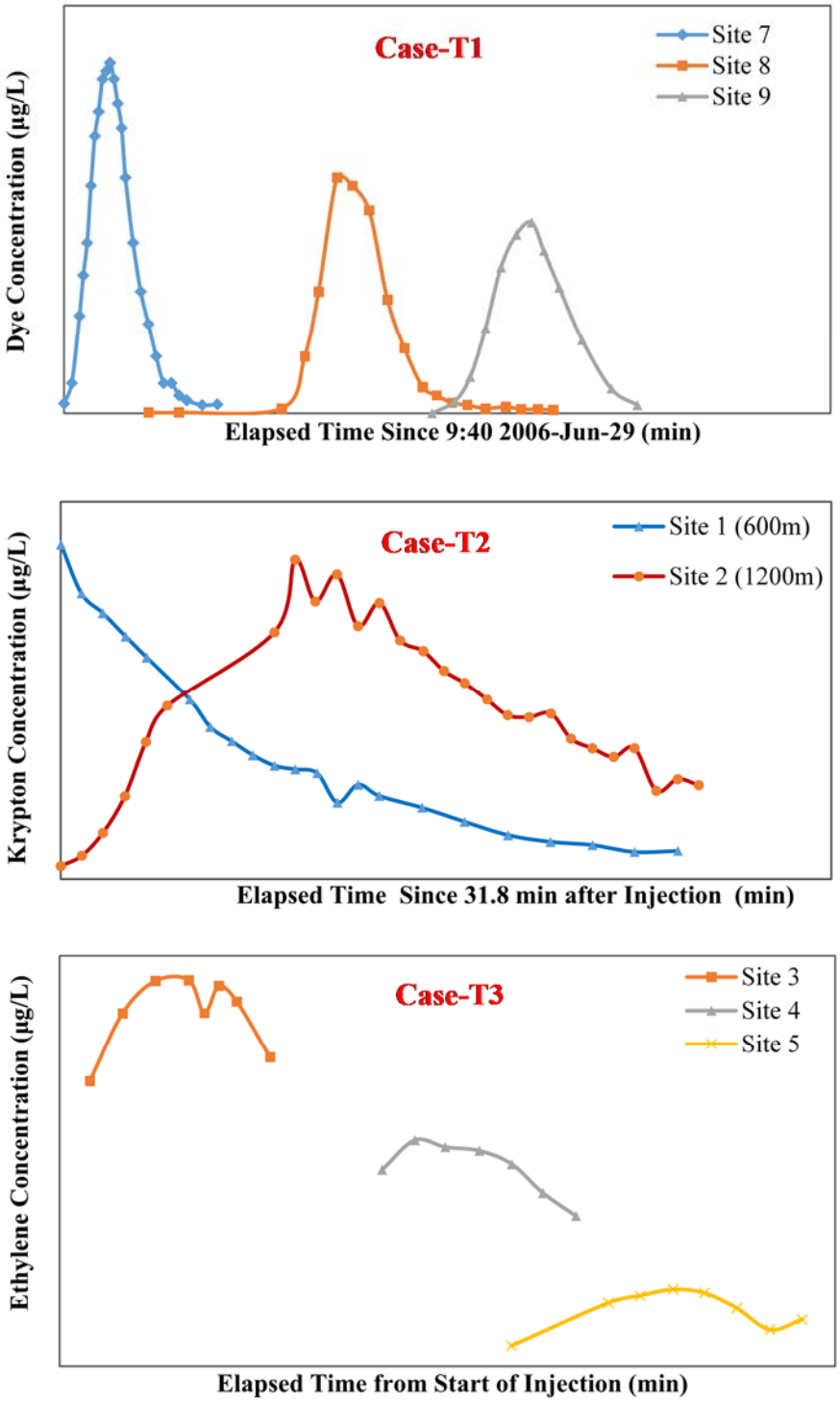
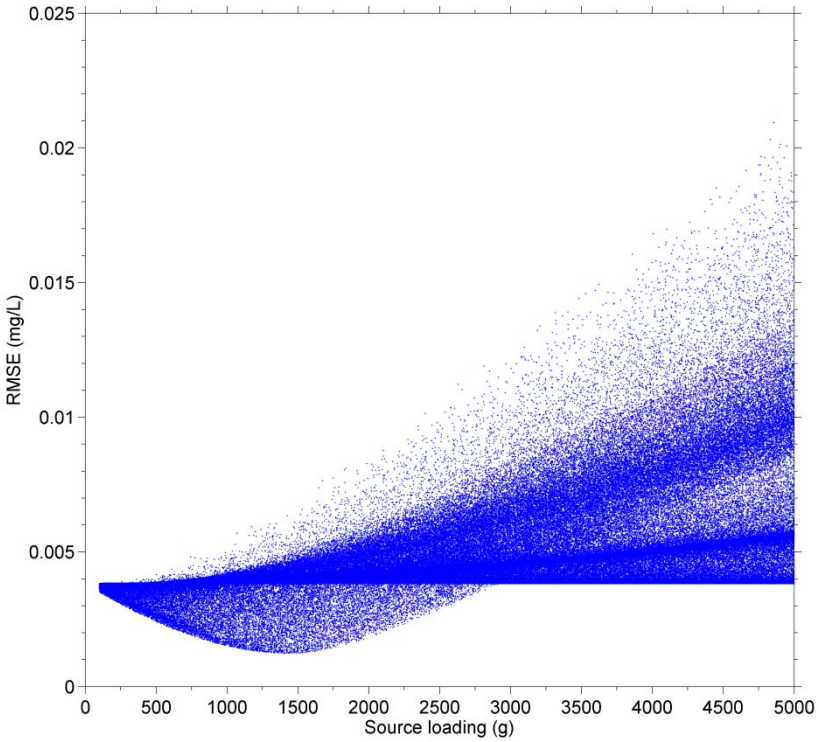
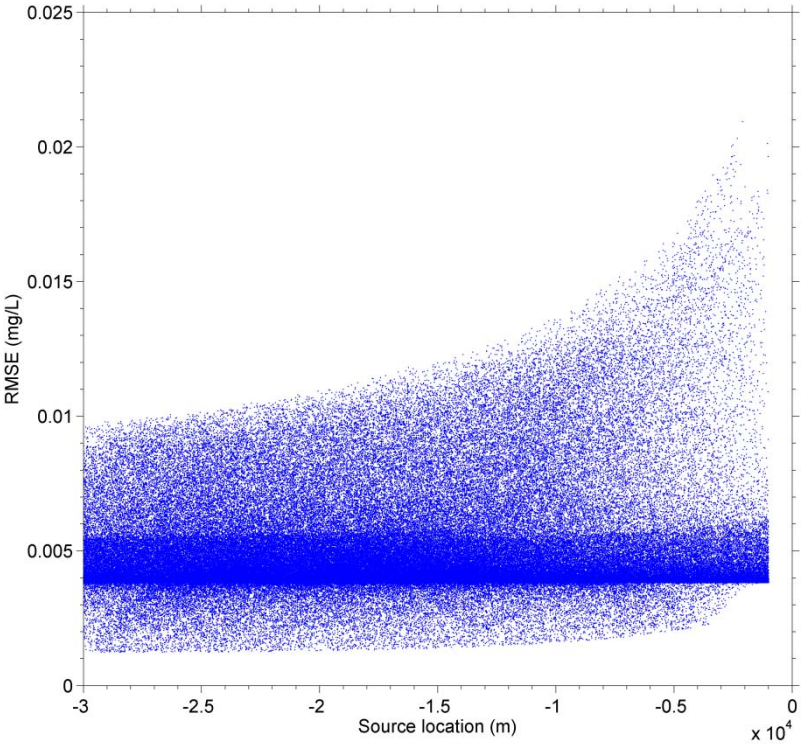


Fig. S9 Tracer concentration data in field tracer experiments


0.025
0.02
0.015
0.01
0.005
0
RMSE (mg/L)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Source loading (g)


0.025
0.02
0.015
0.01
0.005
0
RMSE (mg/L)
-3
-2.5
-2
-1.5
-1
-0.5
0
Source location (m)
x 10^4

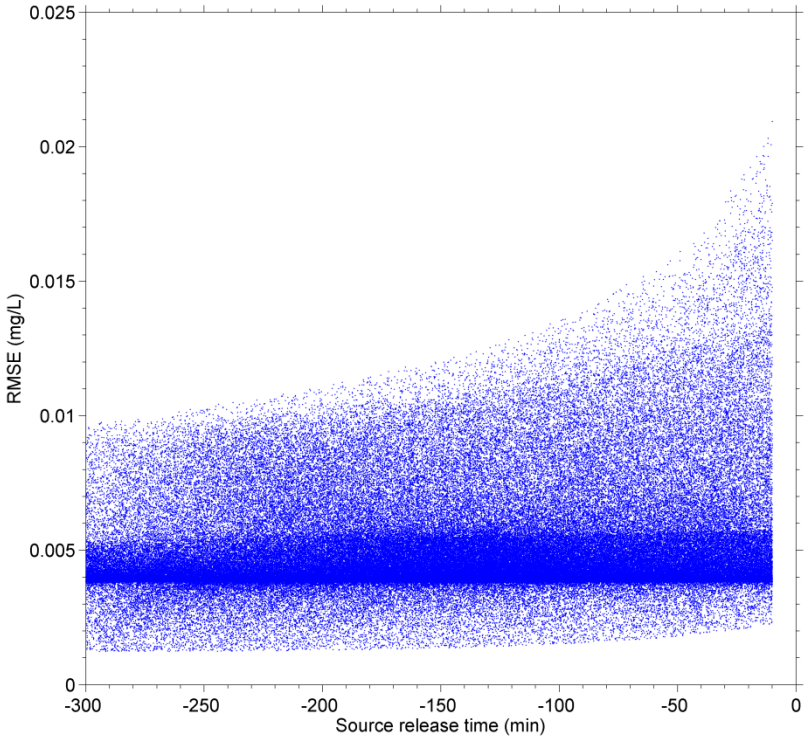


Fig. S10 A typical dotty plots of one run on Case-T1 with 30,000 samples

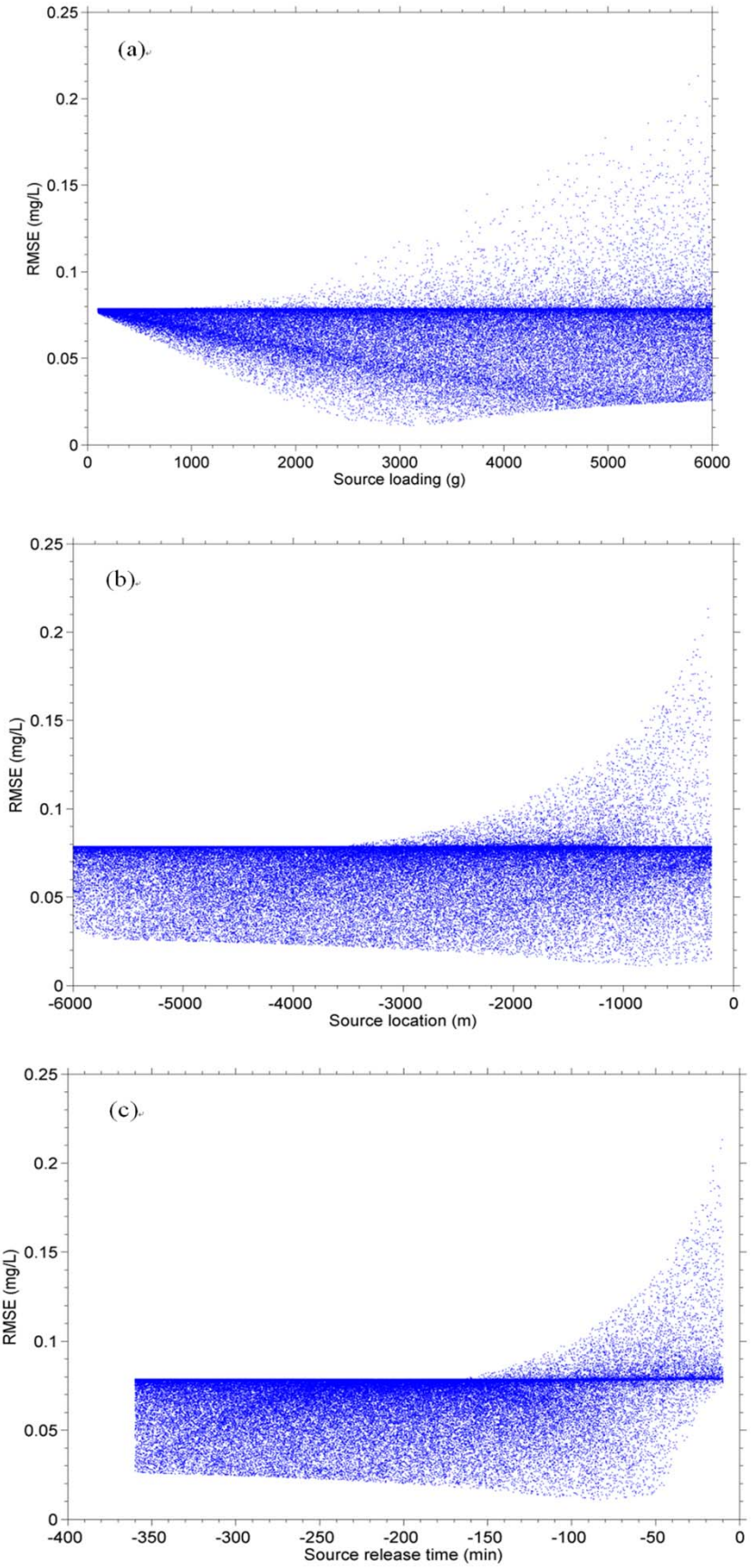


Fig. S11 Dotty plots of RMSE values in Case-S1 (20% noise added) using Direct Monte Carlo. Projection on (a) M_s-RMSEl plane, (b) x_s-RMSE plane, and (c) t_s-RMSE plane.

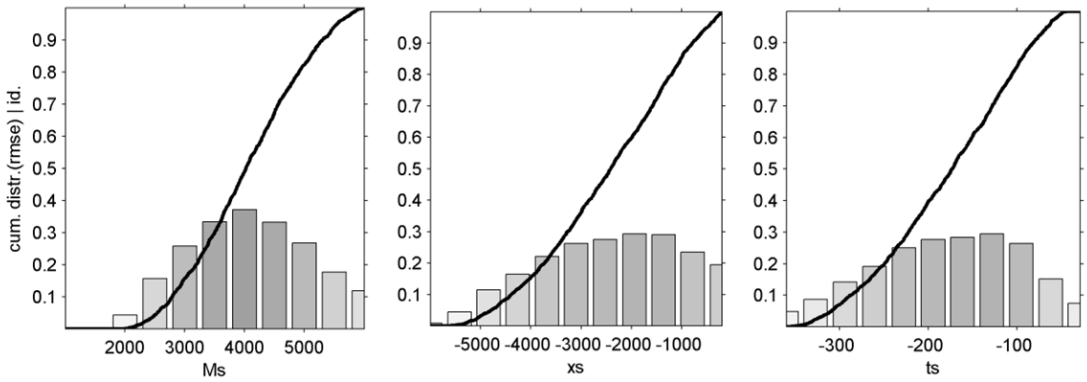
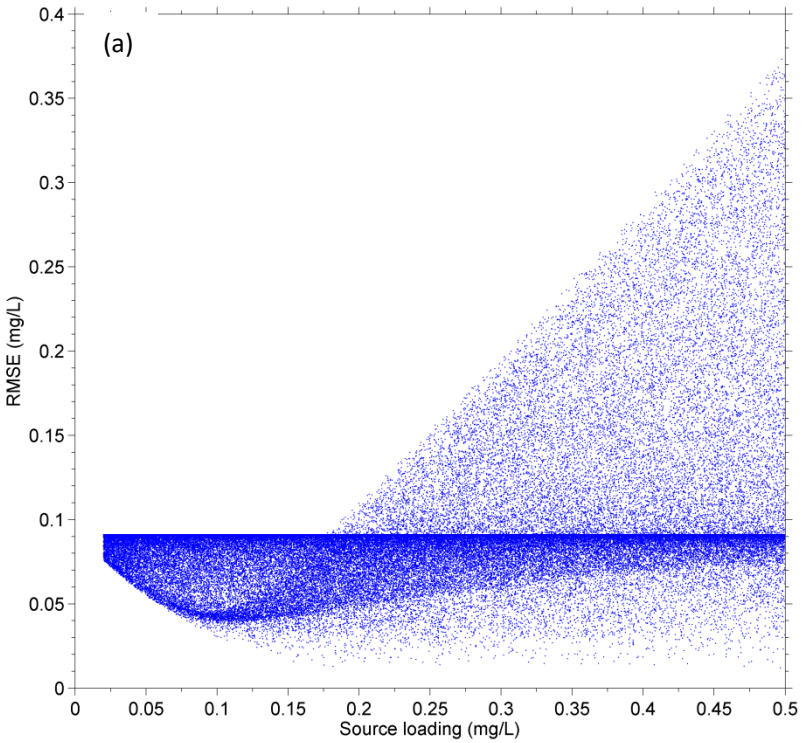
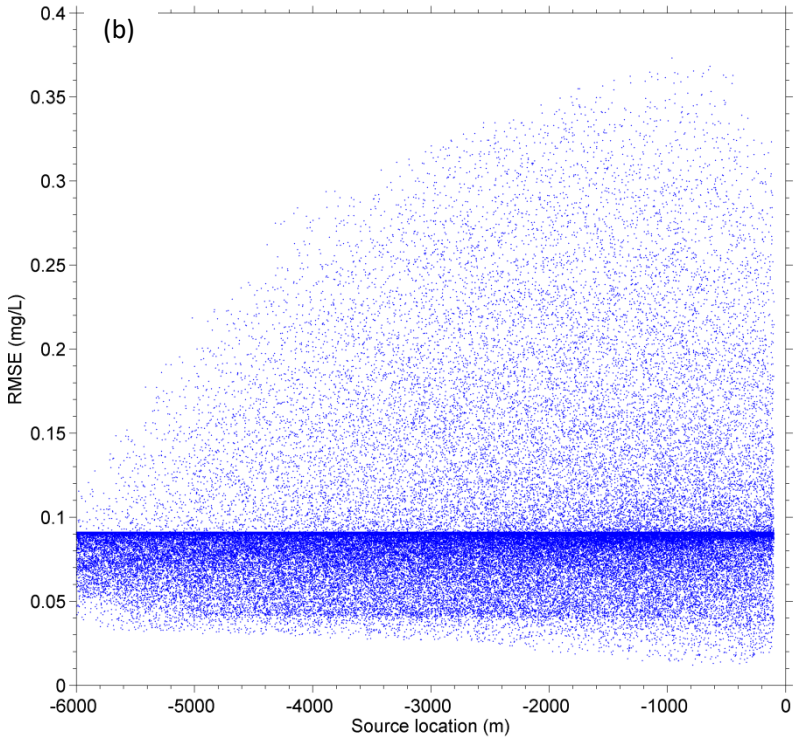


Fig. S12 Identifiability plot of Case-S1 (instantaneous release model). Top 10% of the parameter population in terms of RMSE smaller than 0.05. High gradients in the cumulative distribution indicate high identifiablity in source parameters whereas shallower gradients indicate low identifiablity.


(a)
RMSE (mg/L)
Source loading (mg/L)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5


(b)
RMSE (mg/L)
Source location (m)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-6000
-5000
-4000
-3000
-2000
-1000
0

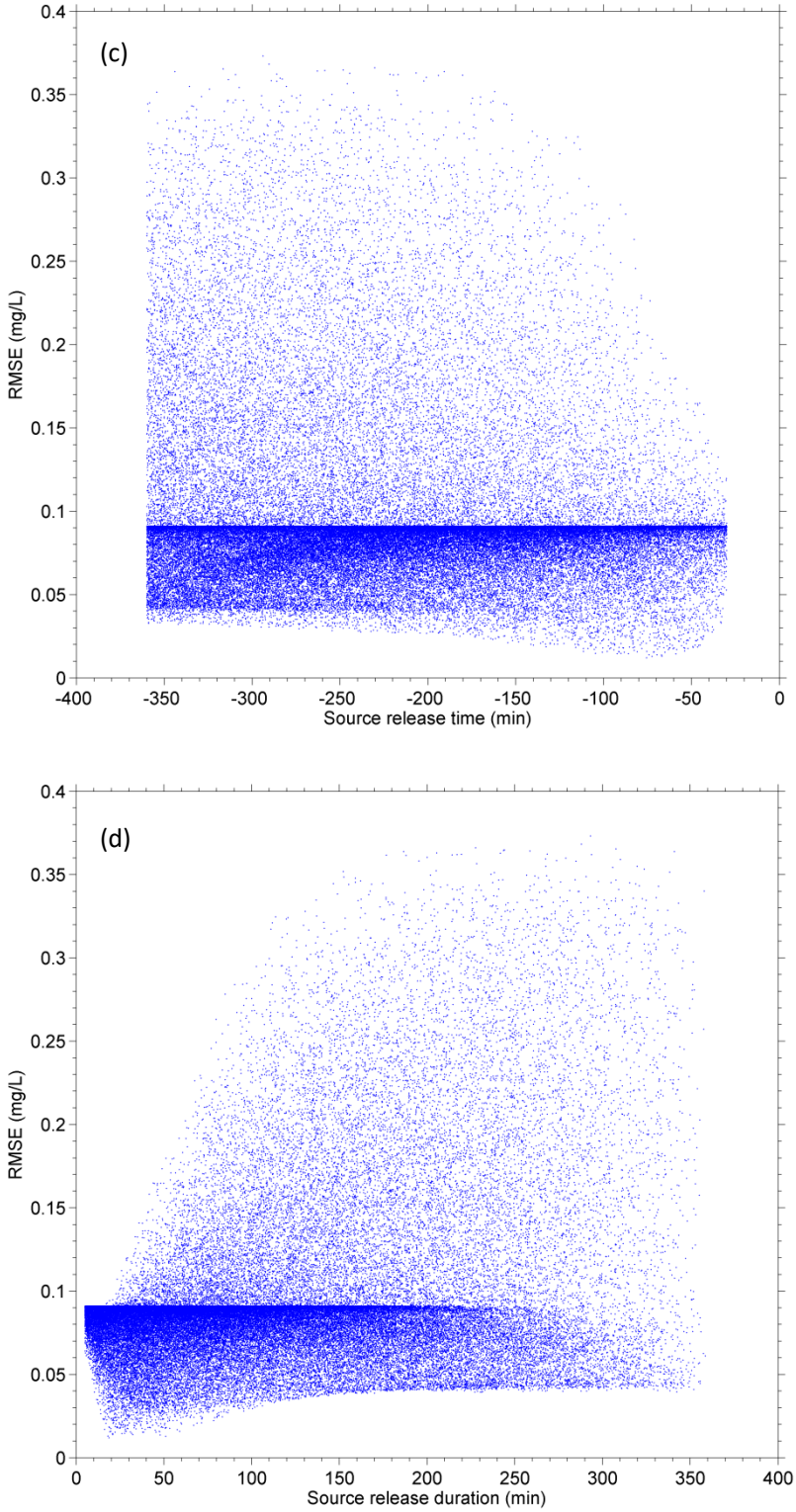


Fig. S13 A typical dotty plot of one run in Case-S2. Projection on (a) C_s-RMSE, (b) x_s-RMSE, (c) t_s-RMSE, and (4) τ_s-RMSE plane. Optimal source parameters and ranges cannot be precisely identified from the figure directly, especially for source loading C_s.

Table S1 Objective functions tested for source inversion in this study

Name	Description	Equation
RMSE	Root mean square error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{fit} - C_i^{obs})^2}$
SQRT- RMSE	Square Root RMSE after data extracted the square root	$SQRT\text{-}RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sqrt{C_i^{fit}} - \sqrt{C_i^{obs}})^2}$
LOG-RMSE	RMSE after data taking normal logarithm	$LOG\text{-}RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\ln(C_i^{fit}+\overline{C_i^{obs}}/10) - \ln(C_i^{obs}+\overline{C_i^{obs}}/10)]^2}$
RSR	RMSE-observations standard deviation ratio	$RSR = \frac{RMSE}{STDEV_{obs}} = \frac{\sqrt{\sum_{i=1}^{N}(C_i^{fit} - C_i^{obs})^2}}{\sqrt{\sum_{i=1}^{N}(C_i^{obs} - \overline{C_i^{obs}})^2}}$
PBIAS	Percent bias	$PBIAS = \frac{\sum_{i=1}^{N}\lvert C_i^{fit} - C_i^{obs}\rvert}{\sum_{i=1}^{N}C_i^{obs}} \times 100$
NSE	Nash-Sutcliffe efficiency	$NSE = 1 - \frac{\sum_{i=1}^{N}(C_i^{fit} - C_i^{obs})^2}{\sum_{i=1}^{N}(C_i^{obs} - \overline{C^{obs}})^2}$
SQRT-NSE	NSE after data taken the square root	$SQRT\text{-}NSE = 1 - \frac{\sum_{i=1}^{N}(\sqrt{C_i^{fit}} - \sqrt{C_i^{obs}})^2}{\sum_{i=1}^{N}(\sqrt{C_i^{obs}} - \sqrt{\overline{C^{obs}}})^2}$
TIC	Theil's inequality coefficient	$TIC = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i^{fit} - C_i^{obs}\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{fit})^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{obs})^2}}$

Where, N is the number of concentration data detected by emergency monitoring; i is the index of a sample point, located at x_i and detected at time t_i; C_i^{fit} and C_i^{obs} are the concentration estimated from fitted model and observed in the field at sample point i, respectively; $\overline{C^{obs}}$ is the mean value of observed concentration data. The word 'error' above denotes the fitting residual or fitting error in this investigation.

As a most commonly used error statistic, RMSE is sensitive to extreme mismatches between C_i^{fit} and C_i^{obs}. Log-RMSE will be more influenced by low concentrations, whereas SQRT-RMSE will represent all the concentration values, giving equal weight to both high and low concentrations (Chahinian et al. 2006). The PBIAS index measures the average tendency of simulated data to be underestimated or overestimated than corresponding observed ones (Gupta et al. 2009). NSE ranges from -∞ to 1 (perfect fit) and determines the relative magnitude of residual variance to original variance (Nash and Sutcliffe 1970; Moriasi et al. 2007). TIC values, or U-statistics, range from 0 (perfect fit) to 1 (maximum inequality) showing that the relationship between a pair of time series data are significantly different or identically same (Bliemel 1973; Murray-Smith 1998). Details for the advantages and disadvantages of those residual measures refer to above mentioned references.

Bliemel F. (1973). Theil's Forecast Accuracy Coefficient: A Clarification. *Journal of Marketing Research,* **10** (4): 444-446.

Chahinian N., Andreassian V., Duan D., Fortin V., Gupta H., Hogue T., Mathevent T., Montanari A., Moretti G., Moussa R., Perrin C., Schaake J., Wagener T. Xie Z. (2006). Compilation of MOPEX 2004 results. In: Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment-MOPEX, 313-338.

Gupta H. V., Kling H., Yilmaz K.K., Martinez G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydroinformatics*, **377** (1-2): 80-91.

Murray-Smith D.J. (1998). Methods for the external validation of contiuous system simulation models: a review. *Mathematical and Computer Modelling of Dynamical Systems*, **4** (1):5-31.

Moriasi D.N., Arnold J.G., Liew M.W.V., Bingner R.L., Harmel R.D., Veith T.L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *T ASABE*, **50** (3): 885-900.

Nash J.E., Sutcliffe J.V. (1970). River flow forecasting through conceptual models part I-A discussion of principles. *Journal of Hydrology*, **10** (3): 282-290.

Results of the tests on synthetic cases

Computing environment

Same as tests based on field experiments cases, the computing environment in synthetic cases is:

- CPU : AMD Dual Core, 3.40GHz
- Memory: DDR2, 3.25GB, 667 MHz
- Video Memory : 256MB (standalone video card)

Inverse instantaneous source (Case-S1)

According to the concentration detected, assuming released pollutant ranges from 100g to 6000g between 10min to 6 hours before the first emergency monitoring taken. Source perhaps locates at -200m to -6000m. 100,000 of source vectors **s** (M_s, x_s, t_s) were produced by identical uniform distributions and objective function are evaluated by RMSE.

Both the projection dotty plots approach and statistic approach show that three source parameters are all successfully estimated by DMC with highly accuracy. Results are reported in Table 3 where results are balanced by 20 runs. Dotty plots are presented in Fig. S11, where optimal source parameters and ranges can be easily obtained from the lowest points. Density denotes how much the overlap of response surface is. The slope near minima reflects the sensitivity of the parameter to the forward model and the robustness or uncertainty of the inversed source parameter. A steep slope is more expected than a mild one.

Identifiable parameters can be characterized if there is a distinct minimum in the dotty plots (Fig. S11). Lack of a distinct minimum indicates the difficulty to find a single optimal value that provides good inversion performance. Hence the parameter is termed poorly identifiable, i.e. more uncertainty for inversion results (Demaria, et al. 2007). Fig. S11 show that pollutant load is most identifiable and source location has the largest uncertainty. Likely, identifiability plots (Wagener and Kollat, 2007) presented at Fig. S12 suggests M_s owns the higher gradients

compared with x_s and t_s, which is consistent with dotty plots.

CPU time consumed from random number (100,000 samples) generating to dotty plots producing is about 42 seconds in average in our computing environment. A first responder would be in favor of this good performance under emergency situation.

Inverse continuous source（Case-S2）

We assume C_s ranges from 0.05 mg/L to 0.5 mg/L, x_s ranges from -6000m to -100 m, t_s ranges from -360 min to -30 min, while τ_s ranges from 5 min to t_s. 10,000 of source vectors **s** (C_s, x_s, t_s, τ_s) were produced by identical uniform distributions and objective function are evaluated by RMSE.

Statistics of inversion results after 20 runs are reported in Table 3. It suggests that DMC is also competent for continuous release scenario. Mean values of each source parameter is quite close to the real values. In term of variability, source location, like in case-S1, and release duration have the largest uncertainty. Standard deviation of estimated source location is about 200m, 30% of mean value. The overall performance of DMC on Case-S1 is better than Case-S2.

Fig.S13 shows the dotty plots of RMSE values on each source parameter in one running. Unlike instantaneously release (Fig. S11), the projection approach cannot capture the precise source parameters directly from the lowest points, especially for source loading, C_s. It partly results from the more complex structure of solution.

In this case, CPU time consumed from random number generating to dotty plot producing is about 114 seconds for 100,000 samples in average in our computing environment. It is slower than Case-S1 because of the constraints from τ_s during random number generation.

Nomenclature

A river cross section area (m^2)

C pollutant concentration (mg/l)

C^{fit} pollutant concentration calculated from forward models (mg/l)

C^{obs} pollutant concentration detected (mg/l)

C_s pollutant concentration in initial cross-section under continuous release scenario

D_x average longitudinal dispersion coefficient (m^2/min*)

D_y average lateral dispersion coefficient (m^2/min)

L_s constant loads (kg/min)

M_s release mass (kg)

Q flow rate (m^3/min)

k first-order decaying coefficient (1/min)

s source term, vector

t monitored time (min)

t_s the initial release time (min)

U,ux longitudinal mean velocity (m/min)

u_y lateral mean velocity (m/min)

x_i location of monitoring sites (m)

x_s release location (m)

x_{mix} distance from source to the place well mixing completed, i.e. length of mixing zone (m)

W river width (m)

τ duration of chemical discharge

Note: take minute (min) as time unit in this work according to the time scale of spill emergency response.